KNN regression experiments

In class we learned about how KNN regression works, and tips for using KNN. For example, we learned that data should be scaled when using KNN, and that extra, useless predictors should not be used with KNN. Are these tips really correct?

In this notebook we run a bunch of tests to see how KNN is affect by the choice of k, distance function, scaling of the predictors, presence of useless predictors, and other things.

One experiment we do not run, and which would be interesting, is to see how KNN performance changes as a function of the size of the training set.

INSTRUCTIONS

Enter code wherever you see # YOUR CODE HERE in code cells, or YOU TEXT HERE in markup cells.

Out[3]: Click here to display/hide the code.

Read the data and take a first look at it

The housing dataset is good for testing KNN because it has many numeric features. See Aurélien Géron's book titled 'Hands-On Machine learning with Scikit-Learn and TensorFlow' for information on the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object 
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

Note that numeric features have different ranges. Fo[redacted]rget [redacted]

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
```

Out[7]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	househol
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.0000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.5396
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.3297
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.0000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.0000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.0000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.0000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.0000

```
# This is formatted as code
```

Missing Data

Notice that 207 houses are missing bedroom info

```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

Out[8]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_inc
290	-122.16	37.77	47.0	1256.0	NaN	570.0	218.0	4.
341	-122.17	37.75	38.0	992.0	NaN	732.0	259.0	1.
538	-122.28	37.78	29.0	5154.0	NaN	3741.0	1273.0	2.
563	-122.24	37.75	45.0	891.0	NaN	384.0	146.0	4.
696	-122.10	37.69	41.0	746.0	NaN	387.0	161.0	3.
...	...	...	...	...	...	...	...	
20267	-119.19	34.20	18.0	3620.0	NaN	3171.0	779.0	3.
20268	-119.18	34.19	19.0	2393.0	NaN	1938.0	762.0	1.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_inc
20372	-118.88	34.17	15.0	4260.0	NaN	1701.0	669.0	5.
20460	-118.75	34.29	17.0	5512.0	NaN	2734.0	814.0	6.
20484	-118.72	34.28	17.0	3051.0	NaN	1705.0	495.0	5.

Let's drop these instances for now

Prepare data for machine learning

We will use KNN regression to predict the price of a house from its features, such as size, age and location.

We use a subset of the data set for our training and test data. Note that we keep an unscaled version of the data for one of the experiments we will run.

```
(7000, 8)
[[ 1.22783551 -1.3492796   0.34639424 -0.16627017  0.11697691 -0.15874461
   0.18687025 -0.74984935]
 [ 0.62095726 -0.82169566  0.58720859 -0.11584049 -0.22077651 -0.0770853
  -0.14171346  1.12877289]
 [-1.16983102  0.7563873  -0.45632025 -0.32112946  0.02736886 -0.37395092
  -0.04890738 -0.10303138]]
```

Baseline performance

For regression problems, our baseline is the "blind" prediction that is just the average value of the target variable. The blind prediction must be calculated using the training data. Calculate and print the test set root mean squared error (test RMSE) using this blind prediction. I have provided a function you can use for RMSE.

```
test, rmse baseline: 112909.3
```

Performance with default hyperparameters

Using the training set, train a KNN regression model using the ScikitLearn KNeighborsRegressor, and report on the test RMSE. The test RMSE is the RMSE computed using the test data set.

When using the KNN algorithm, use algorithm='brute' to get the basic KNN algorithm.

```
test RMSE, default hyperparameters: [illegible]
```

Impact of K

In class we discussed the relationship of the hyperparameter k to overfitting.

I provided code to test KNN on k=1, k=3, k=5, ..., k=29. For each value of k, compute the training RMSE and test RMSE. The training RMSE is the RMSE computed using the training data. Use the 'brute' algorithm, and

Euclidean distance, which is the default. You need to add the get_train_test_rmse() function

```
1  3  5  7  9  11  13  15  17  19  21  23  25  27  29  done

Test RMSE when k = 3: [redacted]
```

Using the training and test RMSE values you got for each value of k, find the k associated with the lowest test RMSE value. Print this k value and the associated lowest test RMSE value. In other words, if you found that k=11 gave the lowest test RMSE, then print the value 11 and the test RMSE value obtained when k=11.

```
best k = [redacted] best test RMSE: [redacted]
```

Plot the test and training RMSE as a function of k, for all the k values you tried.

Comments

In the markup cell below, write about what you learned from your plot. I would expect two or three sentences, but what's most important is that you write something thoughtful.

[redacted]

Impact of noise predictors

In class we heard that the KNN performance goes down if useless "noisy predictors" are present. These are predictor that don't help in making predictions. In this section, run KNN regression by adding one noise predictor to the data, then 2 noise predictors, then three, and then four. For each, compute the training and test RMSE. In every case, use k=10 as the k value and use Euclidean distance as the distance function.

The add_noise_predictor() method makes it easy to add a predictor variable of random values to X_train or

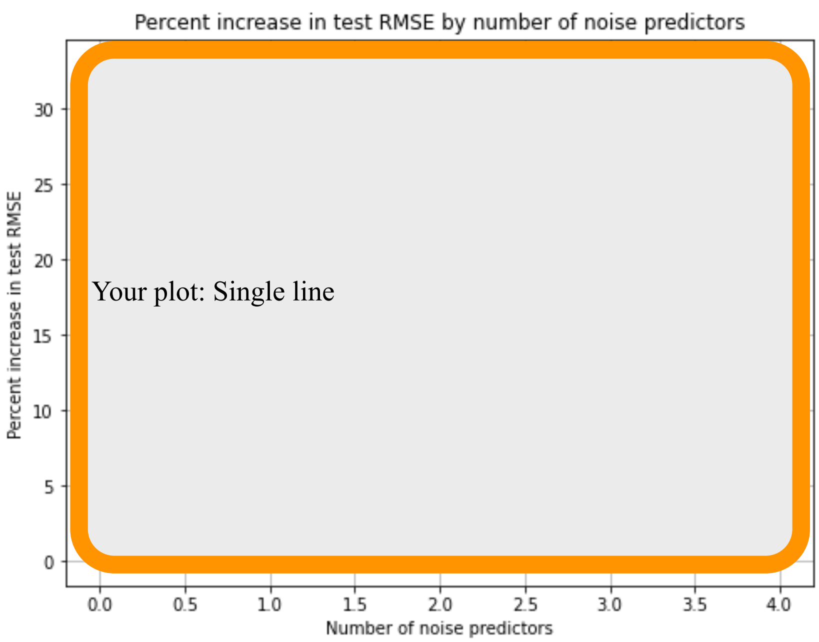
X_test

Hint: In each iteration of your loop, add a noisy predictor to both X_train and X_test. You don't need to worry about rescaling the data, as the new noisy predictor is already scaled. Don't modify X_train and X_test however, as you will be using them again.

```
0  1  2  3  4  done
```

Plot the percent increase in test RMSE as a function of the number of noise predictors. The x axis will range from 0 to 4. The y axis will show a percent increase in test RMSE.

*To compute percent increase in RMSE for n noise predictors, compute 100 * (rmse - base_rmse)/base_rmse, where base_rmse is the test RMSE with no noise predictors, and rmse is the test RMSE when n noise predictors have been added.*



Comments

Look at the results you obtained and add some thoughtful commentary.

[redacted]

Impact of scaling

In class we learned that we should scaled the training data before using KNN. How important is scaling with KNN? Repeat the experiments you ran before (like in the impact of distance metric section), but this time use unscaled data.

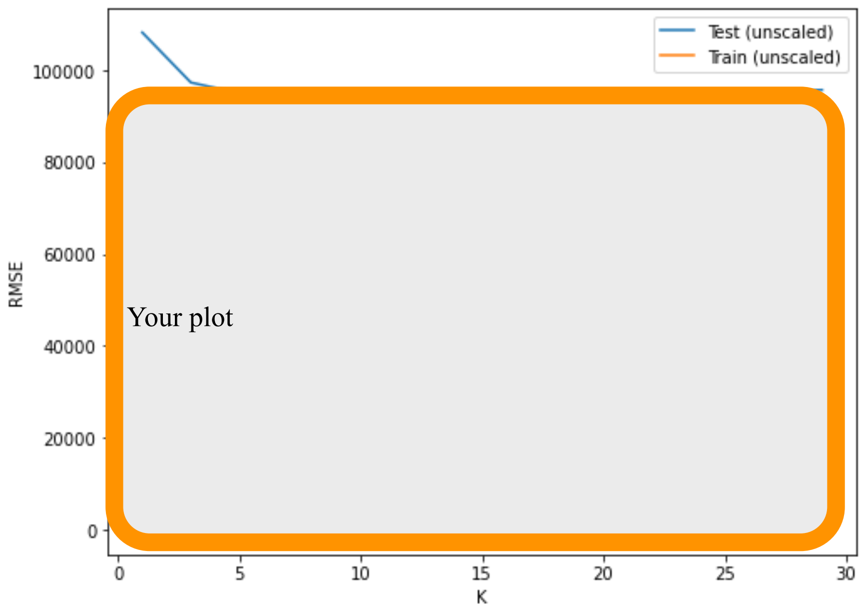
Run KNN as before but use the unscaled version of the data. You will vary k as before. Use algorithm='brute' and Euclidean distance.

```
1  3  5  7  9  11  13  15  17  19  21  23  25  27  29  done
```

Print the best k and the test RMSE associated with the best k.

```
best k = 9, best test RMSE: [redacted]
```

Plot training and test RMSE as a function of k. Your plot title should note the use of unscaled data.



Comments

Reflect on what happened and provide some short commentary, as in previous sections.

[redacted]

Impact of algorithm

We didn't discuss in class that there are variants of the KNN algorithm. The main purpose of the variants is to be faster and to reduce that amount of training data that needs to be stored.

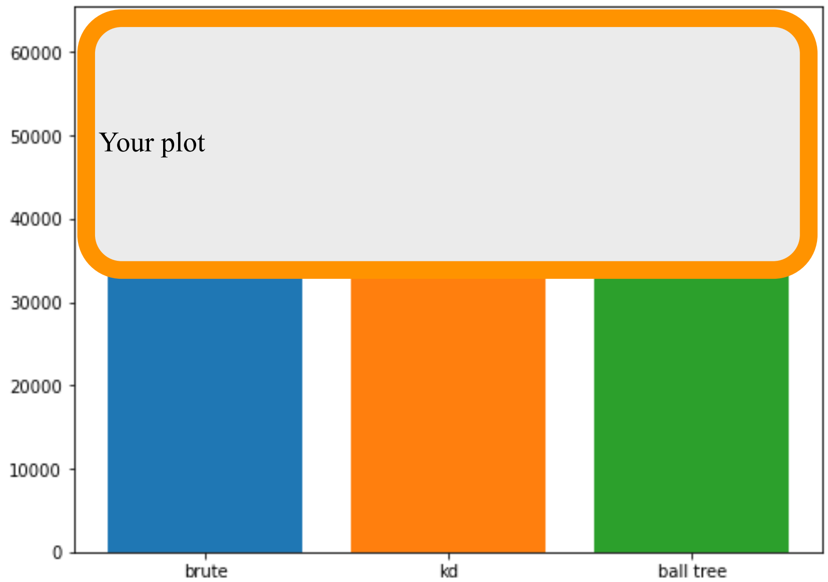
Run experiments where you test each of the three KNN algorithms supported by Scikit-Learn: ball_tree, kd_tree, and brute. In each case, use k=10 and use Euclidean distance.

```
1   3   5   7   9   ball_tree done
1   3   5   7   9   kd_tree done
1   3   5   7   9   brute done
```

Print the name of the best algorith, and the test RMSE achieved with the best algorithm.

```
best ball tree k = 9, best ball tree test RMSE: [redacted]
best kd tree k = 9, best kd tree test RMSE: [redacted]
best brute k = 9, best brute test RMSE: [redacted]
Brute had the best rmse with k=9
```

Plot the test RMSE for each of the three algorithms as a bar plot.



Comments

As usual, reflect on the results and add comments.

[redacted]

Impact of weighting

It was briefly mentioned in lecture that there is a variant of KNN in which training points are given more weight when they are closer to the point for which a prediction is to be made. The 'weight' parameter of KNeighborsRegressor() has two possible values: 'uniform' and 'distance'. Uniform is the basic algorithm.

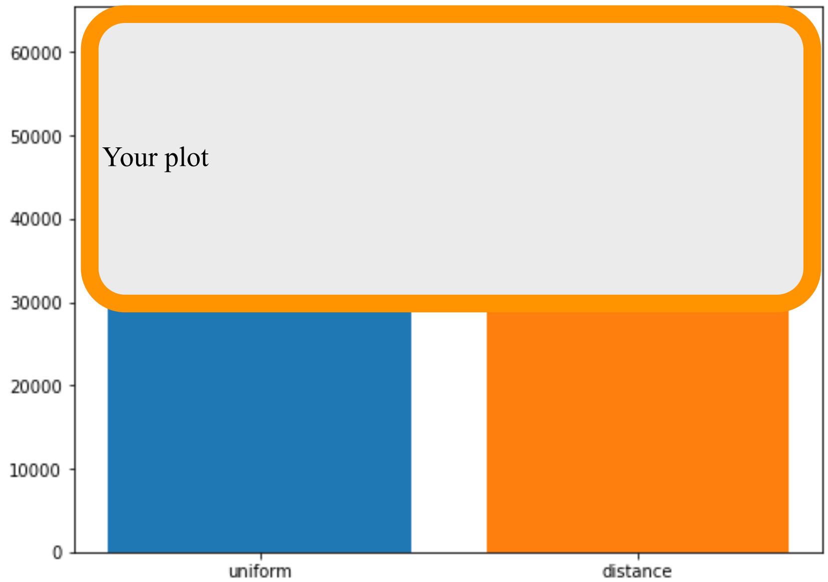
Run an experiment similar to the previous one. Compute the test RMSE for uniform and distance weighting. Using k = 10, the brute algorithm, and Euclidean distance.

```
1   3   5   7   9   uniform weight done
1   3   5   7   9   distance weight done
```

Print the weighting the gave the lowest test RMSE, and the test RMSE it achieved.

```
1  3  5  7  9  uniform weight done
best ball tree k = 9, best ball tree test RMSE: [redacted]
1  3  5  7  9  distance weight done
best ball tree k = 9, best ball tree test RMSE [redacted]
```

Create a bar plot showing the test RMSE for the uniform and distance weighting options.



Comments

As usual, reflect and comment.

[redacted]

Conclusions

[redacted]